Journal of Biomedical Informatics 94 (2019) 103173

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A new data analysis method based on feature linear combination

Xiaohui Lin[a,*], Yanhui Zhang[a], Chao Li[a], Jue Wang[a], Ping Luo[b], Huiwei Zhou[a]

[a] School of Computer Science & Technology, Dalian University of Technology, 116024 Dalian, China
[b] CAS Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, 116023 Dalian, China

ARTICLE INFO

Keywords:
Feature relationship
Classification
Metabolomics

ABSTRACT

In biological data, feature relationships are complex and diverse, they could reflect physiological and pathological changes. Defining simple and efficient classification rules based on feature relationships is helpful for discriminating different conditions and studying disease mechanism. The popular data analysis method, k top scoring pairs (k-TSP), explores the feature relationship by focusing on the difference of the relative level of two features in different groups and classifies samples based on the exploration. To define more efficient classification rules, we propose a new data analysis method based on the linear combination of $k > 0$ top scoring pairs (LC-k-TSP). LC-k-TSP applies support vector machine (SVM) to define the best linear relationship of each feature pair, scores feature pairs by the discriminative abilities of the corresponding linear combinations and selects k disjoint top scoring pairs to construct an ensemble classifier. Experiments on twelve public datasets showed the superiority of LC-k-TSP over k-TSP which evaluates the relationship of every two features in the same way. The experiment also illustrated that LC-k-TSP performed similarly to SVM and random forest (RF) in accuracy rate. LC-k-TSP studies the own unique linear combination for each feature pair and defines simple classification rules, it is easy to explore the biomedical explanation. Finally, we applied LC-k-TSP to analyze the hepatocellular carcinoma (HCC) metabolomics data and define the simple classification rules for discrimination of different liver diseases. It obtained accuracy rates of 89.76% and 89.13% in distinguishing between small HCC and hepatic cirrhosis (CIR) groups as well as between HCC and CIR groups, superior to 87.99% and 80.35% by k-TSP. Hence, defining classification rules based on feature relationships is an effective way to analyze biological data. LC-k-TSP which checks different feature pairs by their corresponding unique best linear relationship has the superiority over k-TSP which checks each pair by the same linear relationship.

Availability and implementation: http://www.402.dicp.ac.cn/download_ok_4.htm.

1. Introduction

In systems biology, building an efficient classifier to discriminate different sample groups is one of the main topics. Simple and powerful classification rules help to interpret the complex physiological and pathological changes as well as study the disease diagnosis and mechanism. Construction of a compact classifier has been vigorously pursued [1]. Popular classification methods, such as support vector machine (SVM), random forest (RF) and neural network (NN) have been widely applied in systems biology [2–7].

To get a good sample classification for analyzing complex biological data, feature selection is usually performed as a preprocessing step before the classifier construction. Many techniques such as SVM-recursive feature elimination (SVM-RFE), Relief and minimal redundancy maximal relevance (mRMR) have been studied and adopted in systems biology [8–10].

Disease is a disordered functioning system of a body that involves multiple factors with the complexity [11]. Exploring changes in feature relationships among different conditions can help to get a deep insight into a multi-factorial basis, which is responsible for the pathogenesis of diseases [12,13].

Maximal information coefficient (MIC) can detect linear and nonlinear relationships between two features [14]. Chen et al. extended MIC to examine the relationship among three features, which is employed to detect pairwise synergy in omics data [15]. Interaction gain measures the interaction of features [16–18], which can tell us whether the two features are independent of each other or they are synergy. The Relative Simplicity (RS) method also evaluates the features by integrating their individual discriminative abilities and their joint effect with others [19].

Chopra et al. studied the feature cooperation in genomics data and constructed the combinatorial variables by *sumdiff, mul* and *sign* [20]. In

* Corresponding author.
E-mail address: datas@dlut.edu.cn (X. Lin).

https://doi.org/10.1016/j.jbi.2019.103173
Received 13 September 2018; Received in revised form 2 April 2019; Accepted 6 April 2019
Available online 06 April 2019

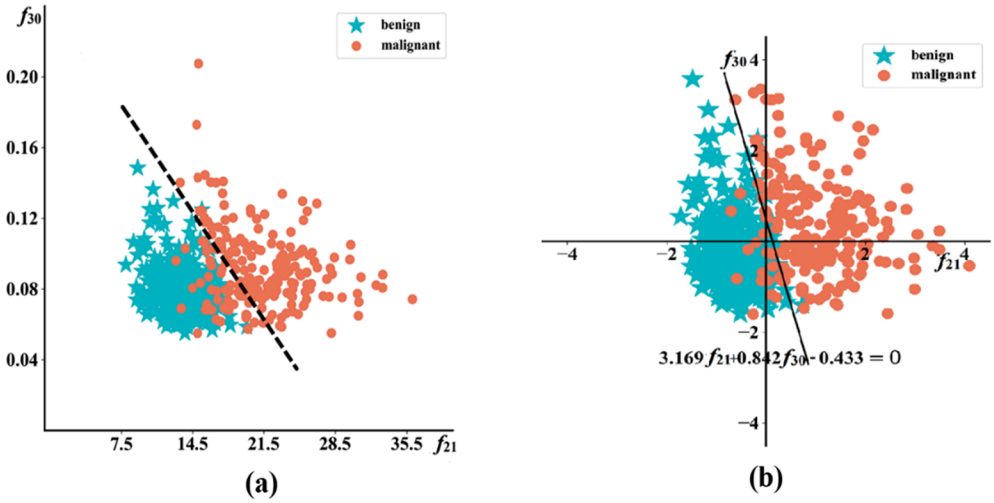

Fig. 1. The scatter plot of samples on the plane of features f_{21} and f_{30} in Wdbc.

addition to the three combinatorial forms, Xing et al. added a new type, *abs*, to further discover the pairwise synergic genes [21].

The methods mentioned above study feature relationships, but they mainly use feature relationships for feature weighting, combination and selection. They do not build the classifiers based on feature relationships. Top scoring pair (TSP) studies relative expression level of every two features, scores each feature pair by the difference of relative expression level of the corresponding two features in different groups, and selects the top scoring feature pair(s) to conduct the classification [22]. The *k*-top scoring pairs (*k*-TSP) is an extension of TSP, it chooses $k > 0$ disjoint top scoring feature pairs and builds an ensemble classifier by the simple majority voting [1]. Based on *k*-TSP, Top Scoring Genes (TSG) fully utilizes the information of sample size to improve the classification performance [23].

k-TSP family algorithms study feature relationships, examine the same relationship of every two features and label the samples based on the feature relationship. They apply several feature pairs to predict the samples, the classification rules are simple and easy to get the biomedical explanation, which is beneficial for studying the nature of pathological changes. But feature relationships in biological data are complex, not all feature pairs reflect pathological changes by the same relationship, such as simple relative expression level in *k*-TSP. Different feature pairs may work in different linear combination forms to reflect the difference between different groups. This paper focuses on the linear combination of features and proposes a method to classify samples based on the linear combination of $k > 0$ top scoring pairs (LC-*k*-TSP). LC-*k*-TSP explores the unique best feature linear relationship by SVM for every two features, it scores each feature pair based on the feature relationship and selects the important feature pairs to define the classification rules. Unlike *k*-TSP which scores each pair by the same linear combination, LC-*k*-TSP scores each feature pair according to its own best linear combination. Experiments on the twelve public datasets and the application in the metabolomics data of liver diseases showed the validity of LC-*k*-TSP.

2. Methods

2.1. TSP and k-TSP

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a dataset with n samples and $x_t \in R^p$ $(1 \le t \le n)$, $F = \{f_1, f_2, \ldots, f_p\}$ be the feature set, $C = \{c_1, c_2\}$ be the class label set. $Y = (y_1, y_2, \ldots, y_n)$ is the class label vector of X and $y_t \in C$ is the class label of sample x_t $(1 \le t \le n)$.

TSP is proposed for binary problems. For feature pair (f_i, f_j) $(1 \le i \ne j \le p)$, TSP examines line $f_i = f_j$ in the plane of f_i and f_j. If most samples of a group lie above line $f_i = f_j$ in the plane, i.e., $f_i > f_j$ in most samples of the group, and most of the other group samples lie below line $f_i = f_j$, i.e., $f_i \le f_j$ in most samples of the other group, then two sample groups could be separated well by line $f_i = f_j$ and TSP assigns a high score to feature pair (f_i, f_j). TSP evaluates each feature pair by this means and selects the pair(s) having the largest score to build a classifier [22].

k-TSP is an extension of TSP, it ranks all feature pairs according to their scores in descending order and selects $k > 0$ top ranked disjoint feature pairs to build an ensemble classifier by majority voting [1].

2.2. LC-k-TSP

Nowadays, genomics, proteomics, and metabolomics, etc. are developing rapidly, they have shown their strong power in disease study and drug development. In omics data, it is known that the relationship among features (genes, proteins, metabolites, etc.) can reflect physiological and pathological changes more robustly than individual features. While omics data are complex, for feature pair (f_i, f_j) $(1 \le i \ne j \le p)$, line $f_i = f_j$ may not be the best separation line in the plane of f_i and f_j. For example, in the public Wisconsin Diagnostic Breast Cancer dataset (Wdbc), which contains 357 benign and 212 malignant samples [24], the simple relative level of features f_{21} and f_{30} cannot separate two group samples (see Fig. 1(a)). The expression level of f_{21} is higher than that of f_{30} in all samples. While, in the plane of f_{21} and f_{30}, it can be seen that the dotted line could separate two sample groups quite well (see Fig. 1(a)). After Z-score standardization, the dotted line in Fig. 1 (a) can be defined as line $3.169\ f_{21} + 0.842\ f_{30} - 0.433 = 0$ (Fig. 1 (b)). It can be clearly seen that most malignant samples locate above line $3.169\ f_{21} + 0.842\ f_{30} - 0.433 = 0$, while most benign samples locate below the line in the plane. Hence, examining the relationship of every two features in the same way of the simple relative expression level is not suitable for all cases. To study feature relationships more comprehensively and define more efficient classification rules, LC-*k*-TSP examines the linear combination of each feature pair, defines the best separation line in the plane of two features and scores the pair by the separation ability of the line.

SVM is a robust supervised learning algorithm. For a subspace of two features f_i, f_j $(1 \le i \ne j \le p)$, SVM could find an optimal hyperplane $\alpha_{ij}f_i + \beta_{ij}f_j + \gamma_{ij} = 0$ by linear kernel, which maximizes the decision boundary to obtain the minimum generalization error [25]. Hence LC-k-TSP adopts SVM to define the best separation line in the plane of two features. For $\alpha_{ij} = 1$, $\beta_{ij} = -1$ and $\gamma_{ij} = 0$, $\alpha_{ij}f_i + \beta_{ij}f_j + \gamma_{ij} = 0$ is just line $f_i = f_j$.

To score feature pair (f_i, f_j), LC-k-TSP defines Δ_{ij} and Γ_{ij} to measure the discriminative ability of line $\alpha_{ij}f_i + \beta_{ij}f_j + \gamma_{ij} = 0$ as follows:

$$\Delta_{ij} = |P_{ij}(c_1) - P_{ij}(c_2)| \tag{1}$$

$$P_{ij}(c_l) = Pr\ (\alpha_{ij}f_i(x_t) + \beta_{ij}f_j(x_t) + \gamma_{ij} > 0 | x_t \in X\ and\ y_t = c_l),\ l = 1, 2 \tag{2}$$

$$\Gamma_{ij} = |u_{ij}(c_1) - u_{ij}(c_2)| \tag{3}$$

$$u_{ij}(c_l) = \frac{\sum_{x_t \in X, y_t = c_l} (\alpha_{ij}f_i(x_t) + \beta_{ij}f_j(x_t) + \gamma_{ij})}{|\{x_t | x_t \in X \quad and \quad y_t = c_l\}| \sqrt{\alpha_{ij}^2 + \beta_{ij}^2}},\ l = 1, 2 \tag{4}$$

$f_i(x_t)$ is the expression value of feature f_i in sample x_t, $P_{ij}(c_l)$ is the probability of the samples in class c_l which lie above line $\alpha_{ij}f_i + \beta_{ij}f_j + \gamma_{ij} = 0$ in the plane of f_i and f_j. Δ_{ij} is in the range of [0, 1]. Obviously, a larger value of Δ_{ij} indicates that the linear relationship between f_i and f_j is more discriminative and line $\alpha_{ij}f_i + \beta_{ij}f_j + \gamma_{ij} = 0$ could separate two sample groups better. LC-k-TSP ranks all feature pairs based on their Δs in descending order. If more than one feature pairs have the same Δ value, Γ_{ij} is introduced to rank them further. Γ_{ij} is the difference between the average distance from samples of group c_1 to line $\alpha_{ij}f_i + \beta_{ij}f_j + \gamma_{ij} = 0$ and the average distance from samples of group c_2 to the line. This means that if the two group samples lie on two different sides of the line and far away from the line, then their separation is much clear. $|\{x_t \mid x_t \in X \text{ and } y_t = c_l\}|$ refers to the total number of samples in class c_l. $\frac{|\alpha_{ij}f_i(x_t) + \beta_{ij}f_j(x_t) + \gamma_{ij}|}{\sqrt{\alpha_{ij}^2 + \beta_{ij}^2}}$ is the distance from sample x_t to line $\alpha_{ij}f_i + \beta_{ij}f_j + \gamma_{ij} = 0$, $u_{ij}(c_l)$ is the average distance from the samples of class c_l $(l = 1,2)$ to line $\alpha_{ij}f_i + \beta_{ij}f_j + \gamma_{ij} = 0$.

LC-k-TSP selects $k > 0$ top scoring feature pairs to conduct the classification. For the selected feature pair (f_i, f_j), the LC-TSP classifier is defined as follows:

If $(P_{ij}(c_1) > P_{ij}(c_2))$

If $(\alpha_{ij}f_i(x_u) + \beta_{ij}f_j(x_u) + \gamma_{ij} > 0)$ $y_u = c_1$;
Else $y_u = c_2$;

Else

If $(\alpha_{ij}f_i(x_u) + \beta_{ij}f_j(x_u) + \gamma_{ij} > 0)$ $y_u = c_2$;
Else $y_u = c_1$;

where x_u is the input sample.

The k pairs selected by LC-k-TSP are disjoint to ensure the diversity of the k base LC-TSP classifiers, and the final prediction is decided by the majority voting of the k base LC-TSP classifiers. Since LC-k-TSP is proposed for binary problems, k is usually odd.

For the setting of k in k-TSP, Afsari et al. proposed an index to measure the performance of k-TSP on different k values and determined an optimal k value in a certain range [26]. Similarly, we also use the index to calculate the performance of LC-k-TSP on a specific k. Let $DS = \{(f_{i1}, f_{j1}), (f_{i2}, f_{j2}), (f_{i3}, f_{j3}), \ldots\}$ be the disjoint feature pair rank list, then (f_{i1}, f_{j1}) is the 1st feature pair which has the largest score. The index of LC-k-TSP is defined as follows:

$$index(k) = \frac{\sum_{r=1}^{k} \Delta_{ir,jr}}{\sqrt{Var\left(\sum_{r=1}^{k} [I(\alpha_{ir,jr}X_{ir} + \beta_{ir,jr}X_{jr} + \gamma_{ir,jr} > 0)] | y = c_1\right) + Var\left(\sum_{r=1}^{k} [I(\alpha_{ir,jr}X_{ir} + \beta_{ir,jr}X_{jr} + \gamma_{ir,jr} > 0)] | y = c_2\right)}} \tag{5}$$

$Var(\cdot)$ denotes the sample variance, $\Delta_{ir,jr}$ is the Δ value of the rth feature pair in the feature pair rank list DS. $I(\cdot)$ is a sign function. $I(\alpha_{ir,jr}X_{ir} + \beta_{ir,jr}X_{jr} + \gamma_{ir,jr} > 0) = [sign(\alpha_{ir,jr}f_{ir}(x_1) + \beta_{ir,jr}f_{jr}(x_1) + \gamma_{ir,jr} > 0)$, $sign(\alpha_{ir,jr}f_{ir}(x_2) + \beta_{ir,jr}f_{jr}(x_2) + \gamma_{ir,jr} > 0)$, …, $sign(\alpha_{ir,jr}f_{ir}(x_n) + \beta_{ir,jr}f_{jr}(x_n) + \gamma_{ir,jr} > 0)]$, and $sign(\alpha_{ir,jr}f_{ir}(x_t) + \beta_{ir,jr}f_{jr}(x_t) + \gamma_{ir,jr} > 0) = 1$ for $\alpha_{ir,jr}f_{ir}(x_t) + \beta_{ir,jr}f_{jr}(x_t) + \gamma_{ir,jr} > 0$, otherwise $sign(\alpha_{ir,jr}f_{ir}(x_t) + \beta_{ir,jr}f_{jr}(x_t) + \gamma_{ir,jr} > 0) = 0$ $(1 \le t \le n)$.

According to Formula (5), we calculate $index(k)$ for odd k and $1 \le k \le kmax$, and select the k value which could induce the largest index value. Here $kmax$ is a predetermined value, which means we search for the best k value in a certain range of $[1, kmax]$.

If a dataset has p features, there are $p(p-1)/2$ feature pairs. For a large p, there are too many pairs to be examined and not all feature pairs can produce efficient classification rules. If two features are independent of each other or they are redundant, their cooperation could not provide more information than the simple summation of the information provided by each of them individually, then LC-k-TSP gives an optional step to neglect these pairs. Interaction gain (IG) studies the interaction between two features. LC-k-TSP uses IG to determine whether the relationship of a feature pair contains more information. For feature f_i, f_j $(1 \le i \ne j \le p)$, the interaction gain $IG(f_i; f_j; C)$ is defined as follows [16,17]:

$$IG(f_i; f_j; C) = I(f_i, f_j; C) - I(f_i; C) - I(f_j; C), \tag{6}$$

where $I(f_i; C)$ is the mutual information of feature f_i and the class C, $I(f_i, f_j; C)$ is the joint mutual information of two features f_i and f_j relative to C.

$$I(f_i; C) = \sum_{f_i^{'} \in f_i} \sum_{c \in C} p(f_i^{'}, c) log \frac{p(f_i^{'}, c)}{p(f_i^{'})p(c)} \tag{7}$$

$$I(f_i, f_j; C) = \sum_{f_i^{'} \in f_i} \sum_{f_j^{'} \in f_j} \sum_{c \in C} p(f_i^{'}, f_j^{'}, c) log \frac{p(f_i^{'}, f_j^{'}, c)}{p(f_i^{'}, f_j^{'})p(c)} \tag{8}$$

$p(f_i^{'})$, $p(f_i^{'}, c)$, $p(f_i^{'}, f_j^{'}, c)$ represent the probability distribution of f_i, the probability distribution of f_i and C, and the joint probability distribution of f_i, f_j and C, respectively.

The value of interaction gain can be positive, zero or negative. $IG(f_i; f_j; C) > 0$ implies that two features f_i and f_j are synergic, their cooperation could produce more information than the simple sum of $I(f_i; C)$ and $I(f_j; C)$. $IG(f_i; f_j; C) = 0$ implies that two features are independent. $IG(f_i; f_j; C) < 0$ implies that two features are redundant [18]. Hence, LC-k-TSP provides an optional step which neglects the feature pairs whose interaction gain is less than or equal to 0 for the dataset having a large p by simply assigning a value of 0 to the Δ and Γ.

Filtering pairs by means of IG is optional. If the dimension of a dataset is small, then the value of $p(p-1)/2$ is not large, there are not too many feature pairs to be examined. LC-k-TSP can study the relationship of every two features effectively (In this study, if the feature number is less than 100, we study the relationship of every two features).

The details of LC-*k*-TSP algorithm is presented as follows:

Algorithm LC-*k*-TSP

Input: The training dataset X ,the class label vector Y, feature set $F = \{f_1, f_2, \ldots, f_p\}$, and *kmax*;
Output: The LC-*k*-TSP classifier and the corresponding feature subset *FS*;
Begin:
For each feature pair (f_i, f_j) $(1 \le i \ne j \le p)$
{
If (IG (f_i; f_j; C) $\le$ 0) //optional
{$\Delta_{ij} = 0$; $\Gamma_{ij} = 0$; continue;}
Construct SVM model based on f_i, f_j and get the line $\alpha_{ij}f_i + \beta_{ij}f_j + \gamma_{ij} = 0$;
Compute Δ_{ij} and Γ_{ij} of (f_i, f_j) according to Formula (1)–(4);
}
Rank all the feature pairs (f_i, f_j) $(1 \le i \ne j \le p)$ according to Δ_{ij} and Γ_{ij} in descending order and get an order list *O*;
DS = {the *kmax* disjoint top scoring pairs in list *O*};
max_index = 1;
Compute *index*(1) based on *DS* according to Formula (5);
max = *index*(1);
For (k = 3; $k \le kmax$; $k = k + 2$)
{
Compute *index*(k) based on *DS* according to Formula (5);
If (*index*(k) > *max*) { *max_index* = k; *max* = *index*(k); }
}
FS = {the *max_index* top scoring pairs in list *DS*};
Build the LC-*k*-TSP classifier;
Return the LC-*k*-TSP classifier and *FS*;
End.

2.3. Experiments

SVM and RF are very popular machine learning techniques, they have been widely applied to analyze omics data [2–5]. Both *k*-TSP and LC-*k*-TSP are classification methods which try to define the simple and effective classification rules based on feature relationships. *k*-TSP examines each pair by the same linear combination, it has been shown that the performance of *k*-TSP is similar as SVM and RF, but it only uses several feature pairs and is easy to explore the biomedical explanation [1,27]. While LC-*k*-TSP explores the unique best linear combination for each feature pair by SVM, and each pair is evaluated by its own combination form. Hence, to validate the performance of LC-*k*-TSP, it was compared with *k*-TSP, SVM and RF on twelve public datasets. Table 1 gives the detailed information of the twelve datasets [24,28–35]. All the datasets are binary problems.

The implementations of *k*-TSP and LC-*k*-TSP were written in R (the following R packages were used for modeling: switchbox [36] and e1071 [37]). R packages e1071 and randomForest [37] were used for training SVM and RF classifiers. A 5-fold cross-validation was run 50 times to obtain the average performance of each method. Five-fold cross-validation divides the data into 5 equal-sized partitions and performs 5 runs. In each run, only one partition is used for testing, and the remaining four ones are used for training, thus each of the 5 partitions is used for testing exactly once. Take LC-*k*-TSP as an example, for each run, 4 of the 5 partitions were used as the training data, LC-*k*-TSP evaluated the feature pairs on the training data, selected the top *k* feature pairs to build the classifier; and the classifier was tested using the remaining one partition which did not take part in the training process. The size of RF was set to 100 (see Table S1 in supplementary information), the linear kernel was applied for SVM and the penalty factor was set as 1 (Table S2 in supplementary information gives the comparison among SVMs with linear kernel, radial basis function kernel and polynomial kernel, which shows SVM with linear kernel performs better than SVM with radial basis function kernel and SVM with polynomial kernel in most cases). For *k*-TSP, *k* was decided as [26] and *kmax* = 9. For LC-*k*-TSP, *k* was decided by Formula (5) and *kmax* = 9.

Subsequently, LC-*k*-TSP was also applied to one metabolomics data about Hepatocellular carcinoma (HCC). HCC is a malignant tumor, precise diagnosis of HCC is very crucial for patients. The samples were got from First Hospital of Jilin University, the Eastern Hepatobiliary Surgery Institute of the Second Military Medicine University (Shanghai, China) and details can be seen in [38]. There were 611 samples in the data, containing 160 normal controls, 126 cirrhosis (CIR) patients and 325 HCC patients. Ninety-two of the 325 HCC samples were diagnosed as early-stage HCC (S-HCC).

Table 1
Twelve public datasets.

Data	No. of features	No. of Samples	Source
BrcaEr	754	146	[28]
Breast	4869	77	[29]
Colon tumor	2000	62	[30]
GSE78775	961	56	[31]
Promoters	57	106	[32]
Sonar	60	208	[24]
Prostate	6033	102	[33]
Ovarian	1536	54	[32]
GSE28700	556	44	[31]
AKI	701	106	[34]
Wdbc	30	569	[24]
Lymphoma	4026	96	[35]

Table 2
Comparison among *k*-TSP, SVM, RF and LC-*k*-TSP (%).

Data	*k*-TSP	SVM	RF	LC-*k*-TSP
BrcaEr	85.21*	81.49*	85.77*	**86.33**
Breast	62.01	62.50	**63.33**	63.07
Colon tumor	**87.26***	82.97	81.00*	83.86
GSE78775	**78.65**	74.92*	75.90*	78.29
Promoters	68.82*	76.92	**88.76***	77.90
Sonar	68.94*	74.53	**82.88***	74.80
Prostate	**91.85***	90.14*	89.84*	90.95
Ovarian	84.89*	**90.87***	90.12*	86.46
GSE28700	**84.57***	76.13	70.25*	74.93
AKI	70.15*	**77.87***	76.31	75.86
Wdbc	88.70*	**97.23***	96.05*	95.39
Lymphoma	87.42*	**94.60***	90.16*	91.00
W/T/L	4/0/8	5/0/7	6/0/6	

Bold: the highest accuracy rates for each dataset.
* *p-value* of *Wilcoxon rank-sum test* between LC-*k*-TSP and the corresponding method less than 0.05.

3. Results and discussion

3.1. Comparison among k-TSP, SVM, RF and LC-k-TSP

Table 2 shows the comparison among *k*-TSP, SVM, RF and LC-*k*-TSP on the twelve binary classification datasets. The number of the datasets where the corresponding method has higher (or equal or lower) accuracy rate than LC-*k*-TSP is defined as W/T/L (win/ tie/loss).

From Table 2, we can see that LC-*k*-TSP is better than *k*-TSP for 8 of the 12 datasets, and it is significantly better than *k*-TSP for 7 datasets. LC-*k*-TSP significantly loses to *k*-TSP in 3 (Colon tumor, Prostate and GSE28700) of the 12 datasets. In the experiment of LC-*k*-TSP, we adopted the optional step to neglect the feature pairs whose cooperation could not provide more information by IG for the datasets having more than 100 features. In most cases, this step could skip many pairs where two features are not interactive. But among the skipped pairs, there may also exist meaningful pairs in some cases.

Both LC-*k*-TSP and *k*-TSP define the classification rules based on feature linear relationships. *k*-TSP considers the same relationship $f_i = f_j$ for each feature pair (f_i, f_j) $(1 \le i \ne j \le p)$ and adopts line $f_i = f_j$ to separate different sample groups. While LC-*k*-TSP explores the unique best separation line $\alpha_{ij}f_i + \beta_{ij}f_j + \gamma_{ij} = 0$ for each pair (f_i, f_j) by SVM. Different pairs may have different linear relationships. The

Table 3
Performance of LC-*k*-TSP and *k*-TSP on the metabolomics data (%).

Sub-problems	*k*-TSP	LC-*k*-TSP
CIR *vs*. S-HCC	87.99*	**89.76**
CIR *vs*. HCC	80.35*	**89.13**
N *vs*. M	**97.31***	95.87

Bold: the highest accuracy rates for each sub-problem.
* *p-value* of *Wilcoxon rank-sum test* between LC-*k*-TSP and *k*-TSP less than 0.05.

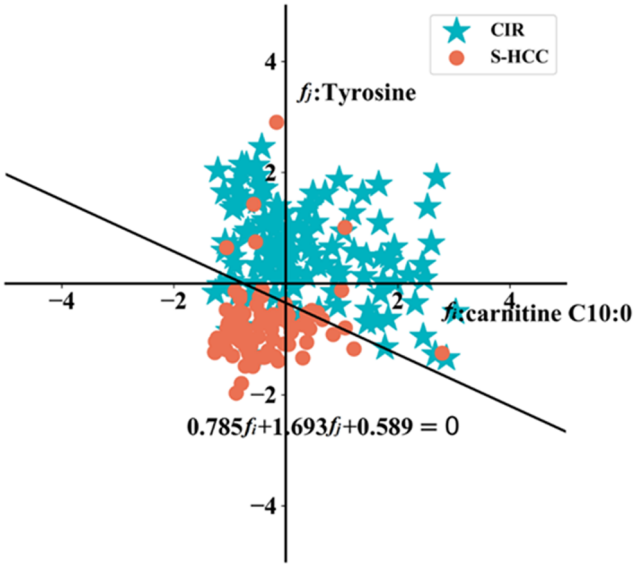

Fig. 2. The scatter plot on feature pair (carnitine C10:0, Tyrosine).

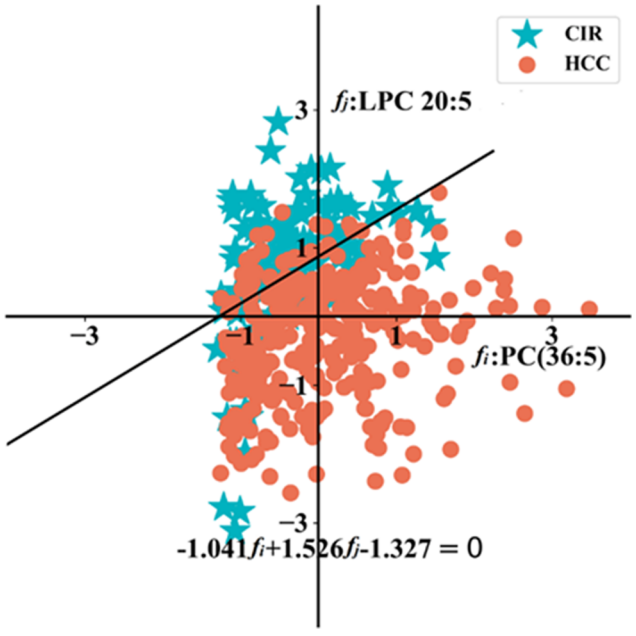

Fig. 3. The scatter plot on feature pair (PC (36:5), LPC (20:5).

experimental results show that exploring its own best linear relationship for each feature pair could better identity different physiological and pathological phenomena than examining each feature pair by the same form as *k*-TSP.

Table 2 also shows that LC-*k*-TSP is superior to SVM for 7 of the 12 datasets, and inferior for 5 datasets. LC-*k*-TSP is better than RF for 6 of the 12 datasets. Both *k*-TSP and LC-*k*-TSP use several feature pairs to label the unknown samples. The performance of *k*-TSP is lower than those of SVM and RF for 7 and 8 of the 12 datasets, respectively. While by defining the unique separation line for each feature pair according to its own situation, LC-*k*-TSP performs similarly to SVM and RF, but it only uses several feature pairs to conduct the classification by the linear combinations, which is simple and easy to study the biomedical explanation.

3.2. Application of LC-k-TSP to the HCC metabolomics data

For the HCC metabolomics study, we aimed at studying the discrimination of different liver diseases, especially studying the discrimination of S-HCC. Hence, we studied the following three binary sub-problems: N (normal controls) *vs*. M (CIR + HCC), CIR *vs*. S-HCC and CIR *vs*. HCC. Table 3 gives the results of LC-*k*-TSP and *k*-TSP on the three sub-problems. A Wilcoxon rank-sum test between the average accuracy rates of LC-*k*-TSP and *k*-TSP has been performed. Usually, it is easy to discriminate normal controls from the diseased. Table 3 shows that both *k*-TSP and LC-*k*-TSP could separate normal controls and diseased quite well, and *k*-TSP is better than LC-*k*-TSP. In reality, it is difficult to discriminate between CIR and HCC, especially CIR and S-HCC. Precise diagnosis of S-HCC can provide early and timely treatment, reduce the death rate of HCC. Table 3 shows that LC-*k*-TSP significantly outperforms *k*-TSP in CIR *vs*. S-HCC and CIR *vs*. HCC. It is true that features (molecules) relate with each other in life activities, but the relationship between different features is different. By exploring the unique linear relationship for each pair, LC-*k*-TSP could separate CIR and S-HCC, CIR and HCC better than *k*-TSP which examines each pair in the same way.

For CIR *vs*. S-HCC and CIR *vs*. HCC, the feature pairs are sorted in descending order according to their frequencies in the 250 runs, the top 9 pairs by *k*-TSP and LC-*k*-TSP were shown in Table S3 and Table S4 in supplementary information, respectively. Now we analyze these frequently selected features in each run of the method and explore the corresponding biomedical meaning. Fig. 2 gives the sample distribution in the plane of the feature pair (f_i, f_j) where f_i is carnitine C10:0 and f_j is Tyrosine (see Table S4 in supplementary information). This pair occurred frequently in the case of LC-*k*-TSP for CIR *vs*. S-HCC, its best separation line is $0.785\ f_i + 1.693\ f_j + 0.589 = 0$. In Fig. 2, the CIR samples mainly locate above line $0.785\ f_i + 1.693\ f_j + 0.589 = 0$, and most of the S-HCC samples locate below the line. This line in the plane of (carnitine C10:0, Tyrosine) separates CIR and S-HCC well. Fig. 3 shows the distributions of CIR and HCC samples in the plane of feature pair (f_i, f_j) where f_i is PC (36:5) and f_j is LPC 20:5 (see Table S4 in supplementary information), this pair occurred frequently in LC-*k*-TSP for CIR *vs*. HCC, the best separation line is $-1.041\ f_i + 1.526\ f_j - 1.327 = 0$. The two group samples could be separate well.

For each of the 9 pairs, a new combinatorial variable is constructed as $f_{ij} = \alpha_{ij} f_i + \beta_{ij} f_j + \gamma_{ij}$ and $f_{ij} = f_i - f_j$ for LC-*k*-TSP and *k*-TSP, respectively, based on the linear relationships defined. Fig. 4 shows the heat maps of the combinatorial variables for CIR *vs* S-HCC. The combinatorial variables based on the 9 feature pairs by LC-*k*-TSP (Fig. 4(b)) are more discriminative than those of *k*-TSP (Fig. 4(a)). Meanwhile, Fig. 5 illustrates the heat maps of the combinatorial variables for CIR *vs*. HCC. It can be seen that the combinatorial variables based on the 9 feature pairs by LC-*k*-TSP could discriminate the different liver diseases very well than those by *k*-TSP.

Table S4 (see supplementary information) gives the detailed information of the 9 feature pairs by LC-*k*-TSP. Among the features in Table S4, GCA and Phe-Trp have been identified in the previous studies [38]. Lipids (LPCs, PCs) play an important role in the regulation of cell invasion, inflammation and cell proliferation [39,40]. Bile acids are important signal metabolites of energy homeostasis [41], the abnormal alteration in the HCC patients may be the result of disturbed energy metabolism. The dysregulated aromatic amino acids and its derivations

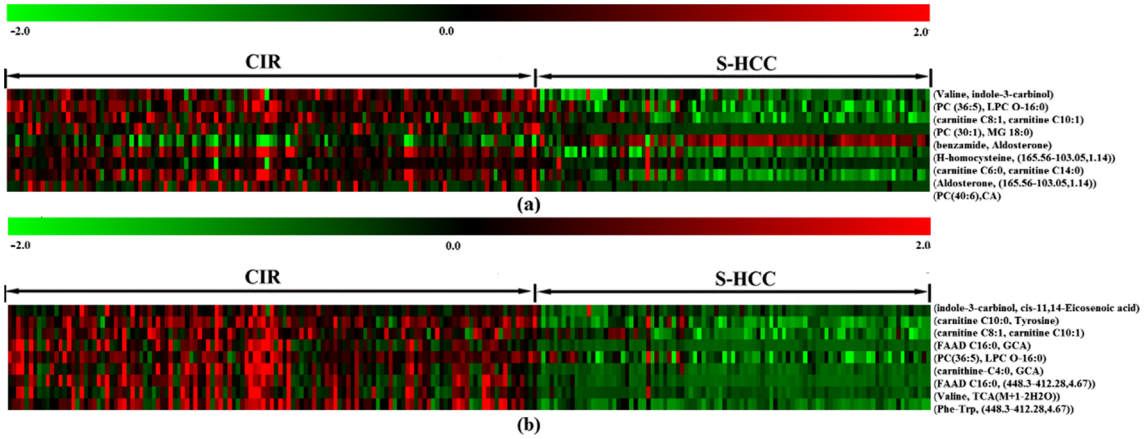

Fig. 4. The heat maps of the combinatorial variables defined by *k*-TSP (a) and LC-*k*-TSP (b) in CIR *vs.* S-HCC.

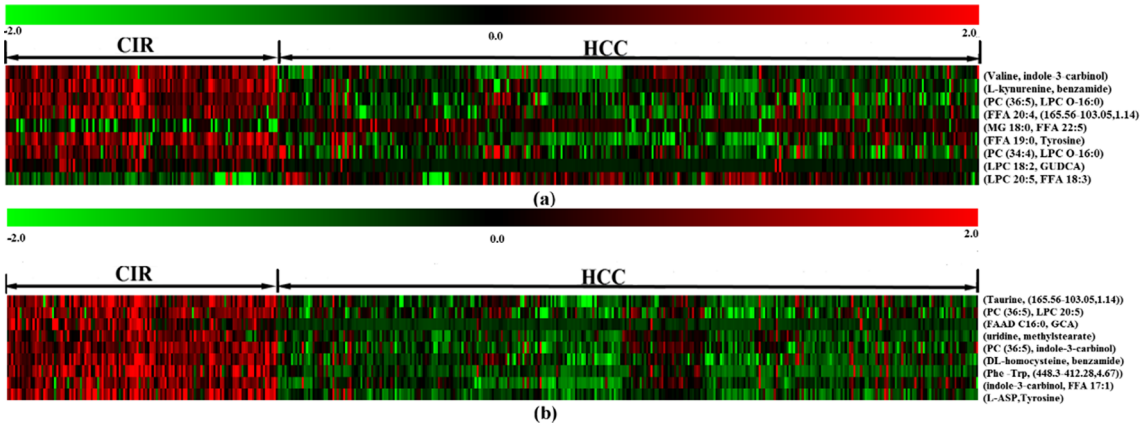

Fig. 5. The heat maps of the combinatorial variables defined by *k*-TSP (a) and LC-*k*-TSP (b) in CIR *vs.* HCC.

may be altered by gut microbiota during the development of HCC [42]. Furthermore, these results confirm that the features selected by LC-*k*-TSP have biological meaning.

Now there are many data analysis techniques, but no one could defeat others in all datasets. The biological data are very complex. The research has shown that the feature relationships can reflect biological physiology and pathological changes. Studying feature relationships can help to interpret the complex physiological and pathological changes and study the disease diagnosis and mechanism [12,13]. If the linear relationships contain meaningful information, LC-*k*-TSP can analyze the data quite well.

4. Conclusions

Defining simple and efficient classification rules for omics data analysis is significant in disease diagnosis and drug development. This study proposes a new data analysis method, LC-*k*-TSP, based on pair-wise linear comparison. LC-*k*-TSP adopts SVM with linear kernel to explore the linear combination of two features, defines the unique best separation line for every two features and selects *k* pairs to build an ensemble classifier. Each base classifier of LC-*k*-TSP is designed based on the unique linear combinatorial form of the corresponding two features. Classification rules are simple and the sample discrimination by feature linear relationships could eliminate individual differences as *k*-TSP. Experiments on the twelve public datasets and the application in the hepatocellular carcinoma (HCC) metabolomics data showed the validity of LC-*k*-TSP and its superiority over *k*-TSP.

Conflicts of interest

The authors declare that there are no known conflicts of interest.

Acknowledgement

The study has been supported by National Natural Science Foundation of China (21375011).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2019.103173.

References

[1] A.C. Tan, D.Q. Naiman, L. Xu, R.L. Winslow, D. Geman, Simple decision rules for classifying human cancers from gene expression profiles, Bioinformatics 21 (2005) 3896–3904, https://doi.org/10.1093/bioinformatics/bti631.
[2] Y. Liu, Active learning with support vector machine applied to gene expression data for cancer classification, J. Chem. Inf. Comput. Sci. 44 (2004) 1936–1941, https://doi.org/10.1021/ci049810a.
[3] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (2000) 906–914, https://doi.org/10.1093/bioinformatics/16.10.906.
[4] R. Diaz-Uriarte, S.A. de Andres, Gene selection and classification of microarray data using random forest, BMC Bioinf. 7 (2006) 3, https://doi.org/10.1186/1471-2105-7-3.
[5] L. Cappellin, E. Aprea, P. Granitto, A. Romano, F. Gasperi, F. Biasioli, Multiclass methods in the analysis of metabolomic datasets: the example of raspberry cultivar volatile compounds detected by GC-MS and PTR-MS, Food Res. Int. 54 (2013) 1313–1320, https://doi.org/10.1016/j.foodres.2013.02.004.

[6] P.S. Maclin, J. Dempsey, Using an artificial neural network to diagnose hepatic masses, J. Med. Syst. 16 (1992) 215–225, https://doi.org/10.1007/bf01000274.

[7] S. Belciug, F. Gorunescu, Learning a single-hidden layer feedforward neural network using a rank correlation-based strategy with application to high dimensional gene expression and proteomic spectra datasets in cancer detection, J. Biomed. Inform. 83 (2018) 159–166, https://doi.org/10.1016/j.jbi.2018.06.003.

[8] X. Lin, C. Li, Y. Zhang, B. Su, M. Fan, H. Wei, Selecting feature subsets based on SVM-RFE and the overlapping ratio with applications in bioinformatics, Molecules 23 (2018) 52, https://doi.org/10.3390/molecules23010052.

[9] L. Gao, T. Li, L. Yao, F. Wen, Research and application of data mining feature selection based on relief algorithm, Journal of Software. 9 (2014) 515–522, https://doi.org/10.4304/jsw.9.2.515-522.

[10] V. Elyasigomari, D.A. Lee, H.R.C. Screen, M.H. Shaheed, Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckcio optimization algorithm and harmony search for cancer classification, J. Biomed. Inform. 67 (2017) 11–20, https://doi.org/10.1016/j.jbi.2017.01.016.

[11] R. Liu, X. Wang, K. Aihara, L. Chen, Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers, Med. Res. Rev. 34 (2014) 455–478, https://doi.org/10.1002/med.21293.

[12] X. Wang, Role of clinical bioinformatics in the development of network-based biomarkers, J. Clin. Bioinf. 1 (2011) 28, https://doi.org/10.1186/2043-9113-1-28.

[13] L.A. Garraway, E.S. Lander, Lessons from the cancer genome, Cell 153 (2013) 17–37, https://doi.org/10.1016/j.cell.2013.03.002.

[14] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, P.C. Sabeti, Detecting novel associations in large data sets, Science 334 (2011) 1518–1524, https://doi.org/10.1126/science.1205438.

[15] Y. Chen, D. Cao, J. Gao, Z. Yuan, Discovering pair-wise synergies in microarray data, Sci. Rep. 6 (2016) 30672, https://doi.org/10.1038/srep30672.

[16] A. Jakulin, I. Bratko, Testing the significance of attribute interactions, Int. Conf. Mach. Learn. 52 (2004), https://doi.org/10.1145/1015330.1015377.

[17] A. Jakulin, I. Bratko, Analyzing attribute dependencies, Lect. Notes Comput. Sci. 2838 (2003) 229–240, https://doi.org/10.1007/978-3-540-39804-2_22.

[18] Z. Zeng, H. Zhang, R. Zhang, C. Yin, A novel feature selection method considering feature interaction, Pattern Recogn. 48 (2015) 2656–2666, https://doi.org/10.1016/j.patcog.2015.02.025.

[19] Y. Chen, L. Wang, L. Li, H. Zhang, Z. Yuan, Informative gene selection and the direct classification of tumors based on relative simplicity, BMC Bioinf. 17 (2016) 44, https://doi.org/10.1186/s12859-016-0893-0.

[20] P. Chopra, J. Lee, J. Kang, S. Lee, Improving cancer classification accuracy using gene pairs, PLoS ONE 5 (2010) e14305, , https://doi.org/10.1371/journal.pone.0014305.

[21] P. Xing, Y. Chen, J. Gao, L. Bai, Z. Yuan, A fast approach to detect gene-gene synergy, Sci. Rep. 7 (2017) 16437, https://doi.org/10.1038/s41598-017-16748-w.

[22] D. Geman, C. d'Avignon, D.Q. Naiman, R.L. Winslow, Classifying gene expression profiles from pairwise mRNA comparisons, Stat. Appl. Genet. Mol. Biol. 3 (2004) Article19, , https://doi.org/10.2202/1544-6115.1071.

[23] H. Wang, H. Zhang, Z. Dai, M.S. Chen, Z. Yuan, TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection, BMC Med. Genomics. 6 (2013) S3, https://doi.org/10.1186/1755-8794-6-s1-s3.

[24] A. Asuncion, D. Newman, UCI repository of machine learning datasets, http://archive.ics.uci.edu/ml/datasets.html.

[25] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, Proc. Fifth Ann. ACM Workshop Comput. Learn. Theory (1992) 144–152, https://doi.org/10.1145/130385.130401.

[26] B. Afsari, U.M. Braga-Neto, D. Geman, Rank discriminants for predicting phenotypes from RNA expression, Ann. Appl. Stat. 8 (2014) 1469–1491, https://doi.org/10.1214/14-aoas738.

[27] X. Lin, J. Gao, L. Zhou, P. Yin, G. Xu, A modified *k*-TSP algorithm and its application in LC-MS-based metabolomics study of hepatocellular carcinoma and chronic liver diseases, J. Chromatogr. B-Analyt. Technol. Biomed. Life Sci. 966 (2014) 100–108, https://doi.org/10.1016/j.jchromb.2014.05.044.

[28] Y. Hoshida, Nearest template prediction: a single-sample-based flexible class prediction with confidence assessment, PLoS ONE 5 (2010) e15543, https://doi.org/10.1371/journal.pone.0015543.

[29] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, Mao Mao, H.L. Peterse, Karin van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, S.H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (6871) (2002) 530–536, https://doi.org/10.1038/415530a.

[30] Z. Zhu, Y.-S. Ong, M. Dash, Markov blanket-embedded genetic algorithm for gene selection, Pattern Recogn. 40 (2007) 3236–3248, https://doi.org/10.1016/j.patcog.2007.02.007.

[31] National Center for Biotechnology Information(NCBI), http://www.ncbi.nlm.nih.gov/geo.

[32] H. Elghazel, A. Aussem, Unsupervised feature selection with ensemble learning, Mach. Learn. 98 (2015) 157–180, https://doi.org/10.1007/s10994-013-5337-8.

[33] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2002) 203–209, https://doi.org/10.1016/s1535-6108(02)00030-2.

[34] H.U. Zacharias, G. Schley, J. Hochrein, M.S. Klein, C. Koeberle, K.-U. Eckardt, C. Willam, P.J. Oefner, W. Gronwald, Analysis of human urine reveals metabolic changes related to the development of acute kidney injury following cardiac surgery, Metabolomics 9 (2013) 697–707, https://doi.org/10.1007/s11306-012-0479-4.

[35] A. Rakotomamonjy, Variable selection using SVM-based criteria, J. Mach. Learn. Res. 3 (2003) 1357–1370, https://doi.org/10.1162/153244303322753706.

[36] B. Afsari, E.J. Fertig, D. Geman, L. Marchionni, switchBox: an R package for k-top scoring pairs classifier development, Bioinformatics 31 (2015) 273–274, https://doi.org/10.1093/bioinformatics/btu622.

[37] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria. (2015). http://www.R-project.org/.

[38] P. Luo, P. Yin, R. Hua, Y. Tan, Z. Li, G. Qiu, Z. Yin, X. Xie, X. Wang, W. Chen, L. Zhou, X. Wang, Y. Li, H. Chen, L. Gao, X. Lu, T. Wu, H. Wang, J. Niu, G. Xu, A large-scale, multicenter serum metabolite biomarker identification study for the early detection of hepatocellular carcinoma, Hepatology 67 (2018) 662–675, https://doi.org/10.1002/hep.29561.

[39] Y. Tan, P. Yin, L. Tang, W. Xing, Q. Huang, D. Cao, X. Zhao, W. Wang, X. Lu, Z. Xu, H. Wang, G. Xu, Metabolomics study of stepwise hepatocarcinogenesis from the model rats to patients: potential biomarkers effective for small hepatocellular carcinoma diagnosis, Mol. Cell. Proteomics. 11 (M111) (2012) 010694, https://doi.org/10.1074/mcp.M111.010694.

[40] L. Zhou, Q. Wang, P. Yin, W. Xing, Z. Wu, S. Chen, X. Lu, Y. Zhang, X. Lin, G. Xu, Serum metabolomics reveals the deregulation of fatty acids metabolism in hepatocellular carcinoma and chronic liver diseases, Anal. Bioanal. Chem. 403 (2012) 203–213, https://doi.org/10.1007/s00216-012-5782-4.

[41] M. Watanabe, S.M. Houten, C. Mataki, M.A. Christoffolete, B.W. Kim, H. Sato, N. Messaddeq, J.W. Harney, O. Ezaki, T. Kodama, K. Schoonjans, A.C. Bianco, J. Auwerx, Bile acids induce energy expenditure by promoting intracellular thyroid hormone activation, Nature 439 (2006) 484–489, https://doi.org/10.1038/nature04330.

[42] S. Krishnan, N. Alden, K. Lee, Pathways and functions of gut microbiota metabolism impacting host physiology, Curr. Opin. Biotechnol. 36 (2015) 137–145, https://doi.org/10.1016/j.copbio.2015.08.015.